Attorney Docket No. 1331

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant:	Joseph Kevin Gogerty	Date:	February 28, 2003
Serial No.:	09/760,149	Group Art Unit:	1638
Filed:	January 12, 2001	Examiner:	David T. Fox
For:	"INBRED MAIZE LINE PH726"		

Assistant Commissioner for Patents
Washington, D.C. 20231

RULE 132 DECLARATION
OF
DR. STEPHEN SMITH

Sir:

I, Stephen Smith, PhD., do hereby declare and say as follows:

1. I am skilled in the art of the field of the invention. I have a Ph.D. in Biochemical Systematics and Taxonomy of Maize and its Wild Relatives from Birmingham University. I have a M.Sc. in the Conservation and Utilization of Plant Genetic Resources from Birmingham University. I have a Bachelor of Science degree in Plant Sciences from London University. Since 1977 I have been engaged in the development, study and application of molecular markers to genetics, measuring genetic diversity and tracking pedigrees. I commenced this work at North Carolina State University as a post-doctoral research fellow. I have continued my engagement in these studies during my employment by Pioneer Hi-Bred from 1980 until the present. These studies have resulted in numerous scientific articles that have appeared in peer reviewed scientific literature.

2. I have read and understood the Office Action in the above case dated October 30, 2002. This declaration is in response to the Examiner's rejection under, 35 U.S.C. § 112, first paragraph, as containing subject matter which was not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventor(s), at the time the application was filed, had possession of the claimed invention.

3. I have conducted an analysis of Simple Sequence Repeat, SSR, marker data for base inbred PH726 and a backcross conversion of PH726. The trait backcrossed into the backcross conversion of PH726 was male sterility.

Appendix B

4. The SSR data for 457 base inbreds and 103 backcross conversion inbreds, including PH726 and the backcross conversion were used in the analysis. The number of SSR markers for each inbred used in the analysis was between 15 and 87 (mean of 82). The analysis was done as specified in the publication by Berry et al. ("Assessing Probability of Ancestry Using Simple Sequence Repeat Profiles: Applications to Maize Hybrids and Inbreds" Genetics 161:813-824, 2002), with modification as described in Berry et al., (2003); Assessing Probability of Ancestry Using SSR Profiles: Application to maize inbred lines and soybean varieties. Genetics (in review), a copy of which is attached hereto.

5. The results of the analysis indicated that through the use of SSR markers PH726 was identified to be the recurrent parent of the backcross conversion of PH726 over all the other inbreds in the data set. The probability associated with the identification of PH726 as the recurrent parent of the backcross conversion was calculated as 1.00.

6. I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Date: 2-28-03 By: Stephen Smith

Stephen Smith

ASSESSING PROBABILITY OF ANCESTRY USING SIMPLE SEQUENCE REPEAT PROFILES: APPLICATIONS TO MAIZE INBRED LINES AND SOYBEAN VARIETIES

Donald A. Berry,* Jon D. Seltzer,† Chongqing Xie,‡ Deanne L. Wright,‡ Elizabeth S. Jones,‡ Scott Sebastian,‡ J. Stephen C. Smith‡

* Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030

† Medtronic Inc., Minneapolis, MN 55432

‡ Pioneer Hi-Bred International, Johnston, IA 50131

SHORT RUNNING HEAD:

Probability of ancestry using SSR

KEY WORDS:

Inbred alleles, Parentage, Pedigree, SSR, Bayes' Rule

CORRESPONDING AUTHOR:

Donald A. Berry, Ph.D., Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Blvd., Unit 447, Houston TX 77030-4009.

Phone: (713) 794-4141

Fax: (713) 745-4940

E-mail: dberry@mdanderson.org

ABSTRACT

Determining parentage is a fundamental problem in biology and in applications such as identifying pedigrees. Difficulties inferring parentage derive from extensive inbreeding within the population, whether natural or planned; using an insufficient number of hypervariable loci; and from allele mis-matches caused by mutation or by laboratory errors that generate false exclusions. Many studies of parentage have been limited to comparisons of small numbers of specific parent-progeny triplets. There have been few large-scale surveys of candidates in which there is no prior knowledge of parentage. We present an algorithm that determines the probability of parentage in circumstances where there is no prior knowledge of pedigree and which is robust in the face of missing data and mis-typed data. The focus is parentage of an inbred line having uncertain ancestry. The algorithm is a variation of a previously published hybrid-focused algorithm. We describe the algorithm and demonstrate its performance in determining parentage of 43 inbred varieties of soybean that have been profiled using 236 SSR loci and from seven inbred varieties of maize that were profiled using 70 SSR loci. We include simulations of additional levels of missing and mis-typed data to show the algorithm's utility and flexibility.

The determination of parentage using molecular marker data has been little addressed for situations where there is little or no prior knowledge of parentage, or when large-scale surveys involving numerous candidate parents are required. Consequently, we have recently developed an algorithm and demonstrated its use in determining probability of parentage for hybrids in circumstances where there is no prior knowledge of pedigree and which is robust in the face of missing or mis-typed data (Berry *et al.* 2002). We now present a variation of this algorithm that allows determination of parentage for inbred lines or homozygous varieties.

We describe and evaluate a methodology that quantifies the probability of parentage of homozygous genotypes. Our algorithm takes into account that generations of self-pollination occur after the initial parental cross. The number of generations and the initial parental genotypes are unknown. Each generation of inbreeding reduces the number of heterozygous loci in the progeny by an average of 50%. Thus, each of the inbred progeny individuals resulting from the initial parental cross will have lost approximately half of the parental alleles for loci where the inbred parents were fixed for alternate alleles and which were heterozygous in the F1 generation.

The loss of parental alleles during the inbreeding phase is in contrast to the case of a hybrid progeny. An inbred progeny individual will exhibit a lower level of allelic similarity to either of its inbred parents than a hybrid progeny will to its inbred parents. This loss of some parental alleles during inbreeding might be expected to make an inbred algorithm less robust in the face of missing or mis-typed data compared with the hybrid algorithm that has been previously described (Berry *et al.* 2002). We therefore demonstrate the effectiveness and robustness of the inbred algorithm using examples from two species of cultivated plants. We first tested the

algorithm using varieties of the naturally self-pollinating, inbred crop, soybean [*Glycine max. (L.) Merr.*]. This crop was selected because numerous varieties of soybean with known pedigrees were available to us, many of which are closely related. We also used publicly bred inbreds of maize (*Zea mays L.)* that are of known pedigree. Maize is naturally an outcrossing species but inbred lines are most usually generated for use as parents of commercial hybrids. Inbred lines are generated by making successive generations of self-pollination following the initial bi-parental cross.

MATERIALS AND METHODS

Algorithm: The algorithm is a variation of the hybrid version of Berry *et al.* (2002). Consider an index inbred whose parentage is unknown or in dispute. A database containing possible inbred ancestors is available. The objective is to find the probabilities of closest ancestry for each inbred in the database using genotypic information from a large number of SSRs.

Consider a pair of possible ancestors, inbred i and inbred j. We calculate the probability that inbreds i and j are in the index's ancestry, repeating this for all pairs of inbreds in the database. Let $P(i,j|SSRs)$ stand for the posterior probability that i and j are ancestors of the index given the information from the various SSRs. Let $P(i,j)$ stand for the unconditional (or prior) probability of the same event and let $P(SSRs|i,j)$ be the probability of observing the various SSR results if in fact i and j are ancestors of the index. Just as in Berry *et al.* (2002), Bayes' rule relates these various probabilities:

$$P(i,j|SSRs) = P(SSRs|i,j)*P(i,j) / \sum[P(SSRs|u,v)*P(u,v)],$$

where the sum in the denominator is over all pairs of inbreds in the database, indexed by u and v. We need to calculate $P(SSRs|i,j)$ for each i and j. We will make the "no-prior-information" assumption that $P(i,j)$ is the same for all pairs (i,j). Then $P(u,v)$ is a common multiple in the denominator that cancels with $P(i,j)$ in the numerator:

$$P(i,j|SSRs) = P(SSRs|i,j) / \sum P(SSRs|u,v).$$

The problem is to calculate a typical $P(SSRs|i,j)$, the probability of observing the index's SSRs assuming inbreds i and j are both ancestors. The nature of breeding before the self-pollination process is unknown. Since the creation of an inbred proceeds by multiple generations of self-pollination on a hybrid, we label the (unknown) hybrid used to create the (known) index inbred as the intermediate hybrid. When the intermediate hybrid is an immediate descendent of i and j, it receives one of inbred i's alleles and one of inbred j's alleles. When the intermediate hybrid is a second generation descendent of i and j, it receives one allele from each with probability 0.5. And so on. Since degree of ancestry (if any) is unknown, we label the actual probability of passing on one of these alleles to the intermediate hybrid to be p. As in Berry *et al.* (2002) we consider $p = 0.50$ and $p = 0.99$ and here we also consider the intermediate value $p = 0.75$.

When inbreds i and j are ancestors then there are four possibilities: (1) the alleles of both i and j were passed to the intermediate hybrid, (2) i came through but not j, (3) j came through but not i, and (4) neither came through. Assuming independence, these have respective probabilities p^2,

$p(1-p)$, $p(1-p)$, $(1-p)^2$. An allele in the intermediate hybrid's genotype that did not arise from either inbred *i* or inbred *j* is assumed to be selected with probability $1/n$, where n is the total number of alleles at the SSR in question. So far the steps we have described are identical to those for identifying the ancestors of a hybrid described by Berry *et al.* (2002) and, in fact, if the index is heterozygous at an SSR then calculations proceed just as for hybrids. Calculations are substantially different when the index inbred is homozygous, say genotype *aa*. Cases that must be considered are shown in Table 1, where *x* is any allele different from *a* (but not missing). All alleles other than *a* can be grouped because only *a* appears in the index's genotype. For example, *xx* might be *bc* or *bd* or *bb*.

$P(SSR|i,j)$ is the probability of observing the index assuming inbreds *i* and *j* are ancestors. The calculations for SSRs 1 to 6 are shown in Table 2, where the four terms in each case are in order of (1), (2), (3), (4) defined in the previous paragraph. Missing alleles are not considered in the examples above. The number of possibilities is large. Here we consider only the case in which inbred *i* is *aa* and both alleles of inbred *j* are missing. Then

$$P(SSR|i,j) = p^2(1/2+1/2*1/n) + p(1-p)(1/2+1/n*1/2) + p(1-p)(1/n) + (1-p)^2(1/n)$$

Another possibility not considered above is that more than two alleles can be observed for an SSR marker run on individual DNA sample. This can be due to SSR locus duplication, homology due to alloploidy, more than one individual plant being sampled for DNA extraction or cross-contamination. In this case we consider all possible pairings of the observed alleles and calculate using a multiple imputation procedure (Little and Rubin, 1987).

To find the overall $P(SSRs|i,j)$, multiply the individual $P(SSR|i,j)$ over the various SSRs. To determine the probability that any particular inbred, say inbred i, is the closest ancestor of the index, sum $P(SSR|i,v)$ over all inbreds v with $v \neq i$. Call this $P(i|SSRs)$. The maximum of $P(i|SSRs)$ for any inbred i is 1. But since there is one closest ancestor on each side of the family, the sum of $P(i|SSRs)$ over all inbreds i is 2.

SSR data: Soybean DNA was extracted from 490 varieties, all of which were bred in, and are adapted to, the United States. Plant material for DNA extraction was sampled from six plants of each variety. Most of the varieties are proprietary products of Pioneer Hi-Bred International. Several (non-patented) commercial varieties from other breeding companies and some important publicly bred varieties were also included. Procedures for obtaining SSR data from soybean were identical to those described for maize by Berry *et al.* (2002) apart from the following modifications: PCR products with different size ranges and labeled with different fluorochromes were pooled and diluted 1:9 with capillary electrophoresis buffer (Applied Biosystems) then 1:4 with dH20. 1.5ul of pooled DNA were added to 10ul formamide containing the molecular weight size standard 400HD ROX (Applied Biosystems, ROX = 6-carboxy-X-rhodamine). Fragment separation was performed using capillary electrophoresis on an ABI3700 platform (Applied Biosystems), with an injection time of 10 sec at 10,000 V and a run time of 4,000 sec at 7,500 V. Forty-three soybean varieties that had both of their parent varieties also included in the dataset were assigned as index varieties. One to two and occasionally three grandparent varieties of several of the index varieties were also included in the dataset. These varieties collectively

represent a broad array of diversity of soybean germplasm that is currently grown in the United States.

Two hundred and thirty-six publicly available soybean SSR markers (http://soybase.agron.iastate.edu/) were used to demonstrate and evaluate the algorithm. These SSR markers were selected following initial screens on a subset of 24 soybean varieties in which they were tested for amplification and the ability to detect polymorphism. The 236 markers gave good genome coverage and collectively mapped across each of the chromosomal linkage groups of soybean.

All allele scores were made without knowing the identities of the soybean genotypes.

Maize SSR data using 70 loci were previously reported by Senior *et al.* (1998) and were obtained directly from the first author. This publication (Senior *et al.* 1998) cites an array of 94 historically important publicly bred lines that have well known and well established pedigrees. This array of public inbreds includes seven inbreds (A632, A634, Mo17, Pa91, Va35, Va99 and W64A) that each have SSR profiles for their parental lines included in the same dataset. Three of these inbreds were developed from a breeding cross of two unrelated parents. These are: Mo17 which was bred from the cross of C.I. 187-2 x C103; Va99, which was bred from the cross Oh07B x Pa91; and W64A which was bred from the cross of WF9 x C.I. 187-2. Other inbred progeny had more complex pedigrees. One inbred (Va35) was bred from the cross C103 x T8 following an additional cross of T8 as the recurrent parent. Two inbreds (A632 and A634) were bred from the cross Mt42 x B14 following additional crosses of B14 as the recurrent parent.

Pa91 was bred from a complex cross involving four inbreds (WF9 x Oh40B) and (38-11 x L317). These seven progeny inbreds therefore provided an index set of maize inbreds for evaluation of the inbred algorithm.

RESULTS

Data quality: The soybean SSR data that were used to evaluate the algorithm had a mean of 5.5% (range 0-19% loci) missing data per variety. For parent-progeny triplets, there was a mean of 1.1% loci (range 0-5%) where a progeny profile was scored for an allele that was not represented by either of the seed sources that represented the parents. The maize SSR data had a mean of 0.7% missing data (only three genotypes had missing data; these were at elevated levels of 5%, 9%, and 36%). A mean of 6.4% parent/progeny triplets (range 4-7%) had SSR progeny profiles that did not share an allele with either of the seed sources that were available to represent the original parental genotypes.

Probability of ancestry applied to soybean data: Figures 1 and 2 present the probabilities of closest ancestry of the top ranking varieties for each of 43 soybean varieties using data from 236 marker loci at $p = 0.50$ (Fig 1) and at $p = 0.99$ (Fig 2).

When the algorithm was used at $p = 0.5$ with data from all 236 loci (Fig 1), then 24/43 (56%) of index varieties had both parents correctly identified in the top two ranked positions, 12/43 (28%) had one parent correctly placed in one of the top two positions, and 7/43 (16%) had none of the actual parents assigned into the top two ranked positions. Thus, when $p = 0.5$ was used, 60/86

(70%) of actual parental varieties were correctly ranked in the top two positions and 26/86 (30%) were incorrectly placed in lower positions.

When the algorithm was used at $p = 0.75$ with data from all 236 loci (data not shown), 28/43 (65%) of index varieties had both parents correctly identified in the top two ranked positions, 11/43 (26%) had one parent correctly placed in one of the top two positions, and 4/43 (9%) had none of the actual parents assigned into the top two ranked positions. Therefore, when $p = 0.75$ was used, 67/86 (78%) of the actual parental varieties were correctly ranked in the top two positions and 19/86 (22%) were incorrectly placed in lower positions.

When the algorithm was used at $p = 0.99$ with data from all 236 loci (Fig 2), then 33/43 (77%) of actual parental varieties were correctly ranked in the top two positions and 10/86 (23%) had one parent correctly placed; all index varieties had at least one parent ranked in the top two positions when the algorithm was used at $p = 0.99$. With p used at 0.99 then 76/86 (88%) of actual parental varieties were correctly assigned; 10/86 (12%) were incorrectly assigned.

Table 3 presents the rankings, probabilities, and pedigrees of varieties that were incorrectly assigned above a true parent. The largest pedigree class (41% of cases where a non-parent ranked above a true parent) of non-parents ranking higher than parents was for varieties that are derivatives of the parent that was misplaced at a lower ranking. The equal second largest classes (each representing 14% of the cases) were for varieties that were (a) full sibs of the true but misplaced parent and (b) full sibs of a grandparent of the variety for which the pedigree was being tested. Other categories (percent of cases in parentheses) were: multiple backcross versions

of the misplaced parent (7%), a derivative of the variety or which the pedigree was being tested (7%), a half-sib of the true but lower ranked parent (7%), a full sib of the variety for which the pedigree was being tested (3%), and a half-sib of the variety for which the pedigree was being tested (3%). Insufficiently detailed pedigree information is available to categorize one variety (3% of cases) that ranked above the true parent

Robustness: The quality of soybean SSR data as received from the laboratory, in terms of missing data and apparently non-Mendelian parent-progeny triplets, have already been presented. Taking these data as an initial starting point, additional levels of missing and mis-typed data were created by simulations and used to explore robustness of the algorithm.

SSR data for five index soybean varieties were used to determine the robustness of the algorithm. Subsets of data were created that included parameters of reduced numbers of loci, additional levels of missing data, additional levels of mis-typed data, and various combinations of these parameters. Simulated levels of missing and mis-typed data were created with a first pass creating missing data, followed by a second pass creating mis-typed data. Therefore, for example, the maximum level of cumulative error from simulated missing and mis-typed data was from 36 to 40%. Five varieties were chosen to represent a range of diversity in respect of both pedigree and SSR profiles. Four varieties had no parents or grandparents in common and one pair of varieties was related by a common parent. All varieties had parents ranked in the top two positions when the algorithm was run at $p = 0.75$ and $p = 0.99$. This selection of varieties therefore provides a means to establish lower boundaries for both the quantity and quality of SSR data that are required to avoid aberrant results.

Table 4 presents the probability of ancestry of the top five ranked varieties for each of five selected soybean index varieties (93B11, A7986, P9443, S38T8 and Young) when the algorithm is run using different numbers of SSR marker loci (50, 100, 150 and 236) at each of two levels of p (0.5 and 0.99). Using $p = 0.5$, the lowest percentage of parents (60%) that were correctly ranked into the top two positions corresponded to using only 50 SSR. Increasing the number of loci to 100 or 150 or 236 increased the ability to identify the actual parents to about 90%. When p was used at a level of 0.99 all parents were correctly ranked into the top two positions for each of the five varieties when data from as few as 50 SSR loci were used.

Table 5 summarizes other aspects of robustness. Namely, we simulated additional levels of missing, mis-typed and missing plus mis-typed data, beyond those that were inherent in the data as provided by the laboratory. When p was used at a level of 0.5, robustness was generally maintained up to an additional level of 20% simulated missing data, so long as data from 100 or more loci were used. Similarly, robustness was maintained for up to 20% additional mis-typed data so long as data from 100 or more loci were used. Likewise, robustness was maintained with up to 18 to 20% additional levels of data error including both missing and mis-typed data, so long as data from 150 or more loci were used. Using data for all 236 loci provided a higher level of robustness, but even then robustness collapsed when 36 to 40% cumulative additional error from missing and mistyped data were simulated into the analysis. The overall level of correct assignation of parent varieties was higher when p was used at a level of 0.99. All parents then were correctly identified, even when data from only 50 loci were used up to an additional level of 10% missing data. When data from 100 or more loci were used then all parents were correctly

identified with up to 20% additional missing data. Robustness started to decline when the algorithm was applied with 10% additional mis-typed data when data from 150 or fewer SSR loci were used. However, robustness was maintained for up to 20% additional mis-typed data when data from 236 SSR loci were used. When additional levels of both incorrect data were applied then robustness was maintained at levels of up to 10% missing plus 10% mis-typed data so long as data from at least 150 SSR loci were used. Robustness was compromised when additional simulations of 20% missing plus 20% mis-typed data were applied even when data from all 236 SSR loci were used.

We then investigated the relationships of varieties to the index genotype whose pedigree was under examination by rerunning the analysis after both parents of the index genotype had been removed from the analysis. Fifteen varieties that had two or more of their grandparents profiled in the dataset were used for this examination. After removing parents, direct pedigreed derivatives of the index genotype ranked first for P9583, in the first three places for A2943 and in the first six places for P9561. Once all parents and derivatives of the index genotype had been removed from the analysis then the following results were obtained. Predominant classes of varieties ranking in the top five positions were (percent of cases in parentheses): derivatives of the grandparent of the index variety (32%), grandparents of the index variety (16%), derivatives of the parents of the index variety (16%), and half-sibs of the index variety (13%). Grandparents ranked among the first four positions for 10 varieties and were in the first place for five varieties. Great-grandparents ranked within the first seven places for three varieties, and a great-great-grandparent ranked in eighth place for one variety. Other varieties that ranked in the first place were usually closely related to the variety whose pedigree was under examination; full-sibs and

half-sibs were the predominant classes of relatives other than grandparents in the first ranking position after parents and direct derivatives of the variety under examination had been removed.

Probability of ancestry applied to corn data: The seven index inbreds of maize were selected because they represented all of the inbred lines published upon by Senior *et al.* (1998) that had all of their inbred parents also included in the SSR dataset. All of the inbred lines published by Senior *et al.* (1998) have well known and well established pedigrees that are fully provided by those authors.

Table 6 presents probabilities of ancestry for the top five ranked inbreds for each of the seven index inbred lines at two levels of p (0.5 and 0.99). For the three progeny that were bred from single crosses without any subsequent use of one of the parents to make a recurrent cross prior to inbreeding (Mo17, Va99, and W64A) then use of the algorithm at either $p = 0.5$ or at $p = 0.99$ resulted in the parental inbreds being ranked in first and second positions. Use of the algorithm at $p = 0.99$ provided greater discrimination for probabilities of ancestry that were assigned to actual parents compared to highest ranking non-parents. This was most noticeable for the case of inbred Va99 which had a relatively low value when used at $p = 0.5$ for parent 2 (0.5221) compared to parent 1 (0.9999) or to the third ranked inbred (and non-parent), Va22 (0.4252). In contrast, when the program was run at $p = 0.99$ then parent1 and parent2 for Va99 had probabilities of 1 and 0.9855, respectively, with the probability of the third ranked inbred being 0.0131.

For each of the three progeny inbreds that originated from breeding schemes that involved one or more additional crosses of one of their parents, using the algorithm at $p = 0.5$ resulted in

placement of the respective recurrent parent with the highest probability of ancestry. Raising the level of p to 0.99 resulted in both parents (B14 = recurrent parent and MT42 the non-recurrent parent) of the index inbred A632 being ranked in the top two places. Using this level of p also caused a higher ranking (third position) for the non-recurrent parent (MT42) of index inbred A634. Use of p at 0.99 did not cause the non-recurrent parent (C103) of index inbred (Va35) to rank into the top five places.

For the index inbred (Pa91) that was bred from a more complex cross involving four inbred lines, the use of p at 0.5 or at 0.99 resulted in the two parents (WF9 and Oh40B) being ranked in second and third places; highest ranked was inbred Va99 (Va99 is derived from the index inbred Pa91). Neither of the two remaining parents of Pa91 ranked in the top five places.

DISCUSSION

The current widely used North American soybean varieties are founded upon a relatively narrow genetic base of diversity. Gizlice *et al.* (1994) document that the U. S. soybean germplasm base is founded upon 20 plant introductions and that subsequent breeding has made repeated use of related parents. Molecular marker comparisons of elite U. S. soybean varieties compared to a sample of exotic varieties reinforce the conclusion that there is a relative paucity of genetic variation in U. S. soybeans. Narvel *et al.* (2000) have shown that the number of alleles detected among the exotics was 30% greater than among U. S. varieties. Thompson and Nelson (1998) report that very little exotic germplasm has been incorporated into the existing U. S. soybean germplasm base. Examining all pairs of pedigree relationships among the 490 soybean varieties

employed in this study showed that approximately 50% of pairwise relationships are related at the level of half-sib or closer; approximately 10% of pairs are related at the level of full-sib or closer. This set of soybean varieties therefore provides the basis for an extremely rigorous evaluation of the ability of SSR data to distinguish between varieties and of this algorithm to identify pedigrees. Pedigree breeding, including the use of related parents, is also commonly applied in the breeding of maize inbred lines. The set of maize inbreds used here thus also provides a meaningful evaluation of the marker data to discriminate among inbred lines and of the joint ability of the algorithm and of the marker data to allow a determination of inbred pedigrees.

Use of the algorithm at $p = 0.99$ rather than at a lower level improved performance in terms of the percentage of correct assignations of parents and provided a greater statistical differential for probabilities for parents in comparison to the highest ranking non-parents. Use of the algorithm at $p = 0.99$ is more appropriate when it is known that the actual parents of the variety under examination are included among the set of index varieties. If it is not known that the parents are included in the index set then use of the algorithm at $p = 0.5$ is more justified (Berry *et al.* 2002). For the soybean varieties, when p was used at 0.99, then 77% of all varieties that were queried for their parents had both parents correctly identified. Eight-eight percent of soybean parents were correctly identified across 43 index varieties that were queried for their parents. All varieties (with the possible exception of one variety where detailed pedigree information was not available) that ranked above true parents were related either to the mis-ranked parent or to the variety that was being queried for its pedigree. Our previous report of the use of an algorithm to determine hybrid pedigrees (Berry *et al.* 2002) showed a higher level of correct parental

determinations at $p = 0.99$. Many of these soybean varieties have a high degree of pedigree relatedness. However, many of the maize inbred lines that were used in the previously reported study (Berry *et al.* 2002) were also highly related. It is, however, likely to be inherently more challenging to correctly identify parents following cycles of inbreeding because half of the alleles that are segregating in the first generation following the initial breeding cross will be subsequently lost as recurring cycles of self-fertilization occur. Thus, many of the alleles that are present in a hybrid, and which can therefore contribute to the identification of its pedigree, do not remain present in an inbred homozygous progeny.

We examined the pedigrees of soybean index varieties when both parents of the index had been removed from the set of candidate varieties. Direct pedigree descendents with the index variety as one parent then usually ranked higher than other varieties, including varieties that were grandparents or sister varieties of the index variety. When all parents and direct derivatives of the index variety were excluded from the analysis then the predominant classes of varieties ranking in the top five positions were derivatives of the grandparent of the index variety (32%), grandparents of the index variety (16%), derivatives of the parents of the index variety (16%), and half-sibs of the index variety (13%). The SSR data that were available to us did not allow a thorough or very precise assessment of how varieties with different degrees of relatedness would rank as members of the pedigree in the event that the true parents were not present in the database. Nonetheless, when parents were excluded from the analysis then varieties that were very closely related to the index variety ranked highest. Direct descendents dependent for their pedigree upon the index variety, if present, tended to rise above varieties included within other classes of pedigree relationship to the index variety. When varieties directly descended by

pedigree from the index variety were also excluded then a grandparent ranked into first position for 33% of the varieties that were examined. Direct pedigree derivatives of one or more of the parents of the index variety had an equal level of occurrence when parents and derivatives of the index variety were excluded. Further investigations of the identification of grandparents will require a dataset including all grandparents of each index variety and will also require a revised algorithm to take account of pedigree contributions from four varieties as opposed to pairs of varieties which forms the basis of the current inbred algorithm.

For the maize inbred line pedigrees, use of the algorithm either at $p = 0.5$ or at $p = 0.99$ resulted in the correct identification of both parents in all cases where the breeding scheme was an initial cross of two parental lines followed by subsequent cycles of inbreeding (i.e. for the inbreds Mo17, Va99 and W64A). The relatively high level of robustness for results with maize inbreds at $p = 0.5$, in contrast to the results obtained from analyzing soybean data (where 56% of varieties had both parents correctly identified when p = 0.5 was used) could be accounted for by the smaller sample size of maize inbreds and by the lower degree of mean pedigree relatedness amongst this selection of inbred lines in comparison to the soybean varieties. Thus while several inbred lines in this set are closely related, there remain many inbreds that have little or no pedigree relationship (Senior *et al.* 1998).

The inbred algorithm correctly identified both parents of the three maize index inbreds that had been bred from bi-parental crosses that involved equal contributions (by pedigree) from both parents. For the three bi-parental crosses that involved subsequent additional crosses of the recurrent parent (and thus significantly biased contributions by pedigree to the index variety

from the recurrent parent) then use of the algorithm correctly identified each of the recurrent parents. The algorithm was unable to identify the non-recurrent parent in most cases, but this result would be expected because one backcross reduces the expected pedigree contribution of the non-recurrent inbred to 25%. More generations of backcrossing using the recurrent parent then further reduce the expected pedigree contribution of the non-recurrent parent by half at each generation (successively to 12.5%, 6.25%, 3.125%) with the pedigree contribution of the recurrent parent rising accordingly. Since several inbred lines of maize are related by pedigree then it is not surprising that the level of pedigree or SSR similarity of a non-recurrent parent to the index progeny can fall below other inbred lines that are related to the index variety. The algorithm was not able to preferentially identify parents of the inbred line Pa91, which was bred from a complex breeding scheme involving four parents with equal contributions by pedigree. A more suitable algorithm is needed to take account of four way crosses. However, such a need is primarily academic because most breeding crosses in commercial maize breeding, and indeed for most crops, are bi-parental.

These soybean data had a mean of 5.5% missing data per variety and a mean of 1.1% loci where a progeny was scored with an allele that was not also scored in either or both parents. Such apparent non-Mendelian or exclusionary profiles can be due to pollen contamination during inbreeding, cross contamination in the field or laboratory, scoring errors in the laboratory (e.g. scoring +A, predominant stuttering, spectral pull-up, secondary binding sites or polymer spikes), or incorrect pedigrees. Another source of apparent exclusion is through the use of a seed source as a parent that is still heterogeneous due to inbreeding being incomplete. Cycles of inbreeding then continue so that when those seed sources are used in the future as sources for SSR profiling

to represent the parental genotype they will have lost alleles due to inbreeding that have already been passed on to a progeny. Alternately, residual heterozygosity within seed sources can result in low frequencies of heterozygotes or off-type segregants which may, by chance, be sampled in the progeny, but not sampled in the parent. In this study we sampled six plants to represent the variety which may be insufficient to capture alleles existing at low frequencies within the seed source. And even if the allele was sampled, it may not have been detected following PCR amplification due to predominance of the most frequent allele and allelic competition effects. Hall (2002) has also reported the occurrence of apparent non-parental SSR alleles. Mutation can also affect SSR profiles. Vigouroux *et al.* (2002) have estimated mutation rates of 7.7×10^{-4} per generation for dinucleotide SSRs and an upper 95% confidence limit of 5.1×10^{-5} for SSRs with longer repeat units. A level of error or discrepancy in expected SSR profiles are thus inevitable for some, if not all crop plants. We therefore evaluated the robustness of the algorithm and dataset by rerunning the algorithm using datasets that were simulated to have up to 20% additional levels of missing plus 20% mis-typed data beyond the level that was received from the laboratory. The algorithm maintained its initial level of robustness with up to an additional level of 10% both missing and mis-typed data, provided data from at least 100 SSR loci were used. Fewer loci (60) were capable of retaining this degree of robustness in the evaluation of the hybrid pedigree algorithm using maize hybrids (Berry *et al.* 2002). The loss of parental alleles that occurs during the inbreeding process, in contrast to their retention in a hybrid progeny compared to its parents, probably underlies the need to use data from a greater number of loci to maintain robustness for the inbred algorithm as compared to the hybrid algorithm.

It was anticipated that determination of pedigrees following cycles of inbreeding might be more challenging to accomplish than to determine pedigrees of hybrids where the total nuclear genetic contributions of both parents are preserved. Nonetheless, these results show that the algorithm can be used effectively to identifying parents of inbred genotypes. Nearly 90% of soybean parents were identified. This is a set of genotypes which, due to the relatively narrow founder base and subsequent cycles of development through the use of related crosses, provides an extremely rigorous test of the algorithm and of the discriminatory power of the marker data. Supplementary data also show the capability of the algorithm to identify parents of maize inbreds that have been developed in a pedigree system using two parents. Use of this algorithm with currently available codominantly expressed molecular marker data has also been shown to have practical feasibility because of the high degree of robustness that is evident and which extends well beyond the realm of aberrant or unexpected marker data that is encountered. These types of error or unexpected marker data can include laboratory error, sampling effects or the use of different seed sources for the actual parental source compared to a more inbred source that becomes available later to represent the parental genotype. This algorithm has application in a number of fields, including conservation biology, population genetics, and to assist in the protection of intellectual property rights.

LITERATURE CITED

Berry, D. A., J. D. Seltzer, C. Xie, D. L. Wright, and J. S. C. Smith, 2002 Assessing probability of ancestry using simple sequence repeat profiles: applications to maize hybrids and inbreds. Genetics 161: 813-824.

Gizlice, Z., T. E. Carter Jr., and J. W. Burton, 1994 Genetic base for North American public soybean cultivars released between 1947 and 1988. Crop Sci. 34: 1143-1151.

Hall, M.A., 2002 Inbred corn plant 01HF13 and seeds thereof. Patent No. US 6,353,161 B1. U.S. Patent Office, Washington DC.

Little, R. J. A. and D. B. Rubin, 1987 *Statistical Analysis with Missing Data*. J. Wiley & Sons, New York.

Narvel, J. M., W. R. Fehr, W-C Chu, D. Grant, and R. C. Shoemaker, 2000 Simple sequence repeat diversity among soybean plant introductions and elite genotypes. Crop Sci. 40: 1452-1458.

Senior, M. L., J. P. Murphy, M. M. Goodman, and C. W. Stuber, 1998 Utility of SSRs for determining genetic similarities and relationships in maize using an agarose gel system. Crop Sci. 38: 1088-1098.

Thompson, J. A. and R. L. Nelson, 1998 Utilisation of diverse germplasm for soybean yield improvement. Crop Sci. **38**: 1362-1368.

Vigouroux, Y., J. S. Jaqueth, Y. Matsuoka, O. S. Smith, W. D. Beavis, J. S. C. Smith, and J. Doebley, 2002 Rate and pattern of mutation at microsatellite loci in maize. Mol. Biol. Evol. **19**: 1251-1260.

Table 1. Calculations of ancestry for homozygous index inbreds: Cases that must be considered for example of genotype *aa*.

SSR	Index	Inbred *i*	Inbred *j*
1	*aa*	*aa*	*Aa*
2	*aa*	*aa*	*Ax*
3	*aa*	*aa*	*Xx*
4	*aa*	*ax*	*Ax*
5	*aa*	*ax*	*Xx*
6	*aa*	*xx*	*Xx*

x is any allele different from *a*, but not missing

Table 2. Probability of observing the index [$P(SSR|i,j)$] assuming inbreds i and j are ancestors: Calculations for SSRs 1 to 6.

SSR	$P(SSR\|i,j)$
1	$p^2(4/4) + p(1-p)(1/2+1/n*1/2) + p(1-p)(1/2+1/n*1/2) + (1-p)^2(1/n)$
2	$p^2(3/4) + p(1-p)(1/2+1/n*1/2) + p(1-p)(1/2*1/2+1/n*1/2) + (1-p)^2(1/n)$
3	$p^2(2/4) + p(1-p)(1/2+1/n*1/2) + p(1-p)(1/n*1/2) + (1-p)^2(1/n)$
4	$p^2(2/4) + p(1-p)(1/2*1/2+1/n*1/2) + p(1-p)(1/2*1/2 + 1/n*1/2) + (1-p)^2(1/n)$
5	$p^2(1/4) + p(1-p)(1/2*1/2+1/n*1/2) + p(1-p)(1/n*1/2) + (1-p)^2(1/n)$
6	$p^2(0/4) + p(1-p)(1/n*1/2) + p(1-p)(1/n*1/2) + (1-p)^2(1/n)$

The four terms in each case are in order of the four possibilities when inbreds i and j are ancestors: (1) the alleles of both i and j were passed to the intermediate hybrid, (2) i came through but not j, (3) j came through but not i, and (4) neither came through. Missing alleles are not considered.

Table 3. Probabilities of ancestry and pedigree relationships for soybean varieties where both parents did not rank above non-parents.

Case no.	Index variety	Rank	Possible ancestor	Probability
1	*95B97*	1	Parent 2	1
		2	Full sib of parent 1	0.5822
		3	Parent 1	0.4124
2	*A2943*	1	Parent 1	0.9977
		2	Multiple backcross of parent 2	0.7999
		3	Parent 2	0.1999
3	*A4595*	1	Parent 1	1
		2	Derivative of parent 2	0.9956
		3	Multiple backcross of parent 2	0.0034
		4	Derivative of Parent 2	0.0006
		5	Half sib of A4595	0.0004
		6	Parent 2	0.0001
4	*Hark*	1	Parent 1	1
		2	Derivative of parent 2	1
		3	Derivative of parent 2	2.1E-09
		4	Derivative of parent 2	1.4E-09
		5	Derivative of Hark	3.1E-10
		6	Derivative of parent 2	1.1E-13
		7	unknown	3.8E-15
		8	Derivative of parent 2	4.6E-17
		9	Derivative of parent 2	4.7E-21
		10	Parent 2	2.7E-21
5	*Kent*	1	Parent 2	1
		2	Derivative of parent 1	0.9990
		3	Derivative of parent 1	0.0011
		4	Parent 1	3.0E-04
6	*P9583*	1	Parent 1	1
		2	Full sib of P9583	0.8801
		3	Parent 2	0.1199
7	*P9641*	1	Parent 2	1
		2	Derivative of P9641	1
		3	Parent 1	3.7E-06
8	*S30J2*	1	Parent 1	1
		2	Derivative of parent 2	0.9321

		3	Parent 2	0.0679
9	*YB30K01*	1	Parent 2	1
		2	Half sib of parent 1	1
		3	Full sib of parent 2	7.9E-09
		4	Half sib of parent 2	3.3E-09
		5	Full sib of grandparent	1.2E-10
		6	Derivative of parent 1	3.0E-11
		7	Full sib of parent 2	2.0E-11
		8	Full sib of grandparent	8.7E-12
		9	Parent 1	1.1E-12
10	*YB41Q01*	1	Parent 2	1
		2	Full sib of parent 1	1
		3	Full sib of grandparent	7.3E-05
		4	Full sib of grandparent	4.1E-09
		5	Parent 1	9.1E-10

Results for 33 (77%) varieties where both parents were ranked first and second are not included in this table (see Figures 1 and 2).

Table 4. Probability of ancestry for five individual soybean varieties using SSR data obtained from different numbers of loci (50, 100, 150, 236).

	L50		L100		L150		L236	
Inbred	Possible ancestor	Prob	Possible ancestor	Prob	Possible ancestor	Prob	Possible ancestor	Prob
P=0.5								
93B11	*XB31C*	0.9461	*XB31C*	1	*XB31C*	1	*XB31C*	1
	A3415	0.8006	*A3415*	0.9362	*A3415*	0.9146	*A3415*	0.9954
	XB38A01	0.0256	*WILLIAMS*	0.0429	*WILLIAMS*	0.0809	*WILLIAMS*	0.0046
	P9271	0.0251	*A3242*	0.0155	*YB30L01*	0.0034	*A3242*	0
	YB30L01	0.0232	*YB30L01*	0.0015	*A3242*	0.0006	*DOUGLAS*	0
A7986	*COOK*	0.7748	*BRAXTON*	0.9725	*BRAXTON*	1	*BRAXTON*	1
	XB63D00	0.2841	*YOUNG*	0.5302	*YOUNG*	0.8910	*YOUNG*	0.9929
	S6262	0.1826	*COOK*	0.3872	*P9641*	0.0404	*XB63D00*	0.0071
	YOUNG	0.1755	*XB63D00*	0.0496	*XB63D00*	0.0254	*96B32*	0
	BRAXTON	0.1065	*P9641*	0.0328	*COOK*	0.0245	*P9641*	0
P9443	*DOUGLAS*	0.8086	*A3415*	0.5557	*FAYETTE*	0.8760	*FAYETTE*	0.9885
	A3415	0.7629	*FAYETTE*	0.4957	*A3415*	0.7034	*DOUGLAS*	0.8847
	WILLIAMS	0.0887	*DOUGLAS*	0.4855	*CX399*	0.1671	*A3415*	0.0846
	YALE	0.0501	*CX260C*	0.2032	*CX260C*	0.1273	*WILLIAMS*	0.0348
	P9394	0.0411	*WILLIAMS*	0.1608	*WILLIAMS*	0.0948	*CX399*	0.0062
S38T8	*S3535*	0.8711	*S3535*	0.9993	*S3535*	1	*S3535*	1
	S4644	0.4543	*S4644*	0.9988	*S4644*	1	*S4644*	1
	YB44R01	0.2762	*YB40M01*	0.0012	*YB37Y00*	0	*A4268*	0
	YB40M01	0.1087	*YB44R01*	0.0004	*93B65*	0	*YB44R01*	0
	YB44Q01	0.0325	*YB37Y00*	0.0001	*A4268*	0	*YB37Y00*	0
YOUNG	*DAVIS*	0.6589	*DAVIS*	0.6551	*DAVIS*	0.6324	*DAVIS*	0.9752
	XB63D00	0.4942	*ESSEX*	0.5079	*P9641*	0.5524	*P9641*	0.5397
	96B32	0.3122	*P9641*	0.3409	*COOK*	0.3231	*ESSEX*	0.3273

	COOK	0.0707	*COOK*	0.1692	*ESSEX*	0.2817	*96B32*	0.1299
	OGDEN	0.0606	*96B32*	0.1315	*96B32*	0.1933	*COOK*	0.0235
p=0.99								
93B11	*XB31C*	1	*XB31C*	1	*XB31C*	1	*XB31C*	1
	A3415	0.9999	*A3415*	0.9999	*A3415*	1	*A3415*	1
	A3242	0.0001	*A3242*	0.0001	*P9443*	0	*WILLIAMS*	0
	P9443	0	*P9443*	0	*A3242*	0	*A3242*	0
	WILLIAMS	0	*WILLIAMS*	0	*WILLIAMS*	0	*FAYETTE*	0
A7986	*BRAXTON*	1	*BRAXTON*	1	*BRAXTON*	1	*BRAXTON*	1
	YOUNG	0.9903	*YOUNG*	0.9903	*YOUNG*	0.9987	*YOUNG*	1
	P9641	0.0092	*P9641*	0.0092	*96B32*	0.0012	*XB63D00*	0
	96B32	0.0005	*96B32*	0.0005	*P9641*	0.0002	*96B32*	0
	DAVIS	0	*DAVIS*	0	*DAVIS*	0	*P9641*	0
P9443	*DOUGLAS*	0.9998	*DOUGLAS*	0.9999	*FAYETTE*	0.9995	*DOUGLAS*	1
	FAYETTE	0.7010	*FAYETTE*	0.7011	*DOUGLAS*	0.9993	*FAYETTE*	1
	CX260C	0.2345	*CX260C*	0.2345	*CX399*	0.0006	*CX260C*	0
	A3415	0.0644	*A3415*	0.0643	*A3415*	0.0005	*CX399*	0
	S3941	0.0001	*AP3330*	0.0001	*P9394*	0.0001	*A3415*	0
S38T8	*S3535*	1	*S3535*	1	*S3535*	1	*S3535*	1
	S4644	1	*S4644*	1	*S4644*	1	*S4644*	1
	YB40M01	0	*YB40M01*	0	*93B67*	0	*A4268*	0
	YB44R01	0	*A5979*	0	*ST3780*	0	*YB54J00*	0
	93B67	0	*YB44R01*	0	*YB37Y00*	0	*YB44R01*	0
YOUNG	*DAVIS*	1	*DAVIS*	1	*DAVIS*	1	*DAVIS*	1
	ESSEX	1	*ESSEX*	1	*ESSEX*	1	*ESSEX*	1
	P9641	0	*P9641*	0	*COOK*	0	*S4240*	0